

Understanding the 2020 Census Disclosure Avoidance System:

Differential Privacy 201 and the TopDown Algorithm

Michael Hawes and Michael Ratcliffe
U.S. Census Bureau

May 13, 2021

Shape
your future
START HERE >

United States[®]
Census
2020

Acknowledgements

This presentation includes work by the Census Bureau's 2020 Disclosure Avoidance System development team, Census Bureau colleagues, and our collaborators, from the following Census Bureau divisions and outside organizations: ADCOM, ADDC, ADRM, CED, CEDDA, CEDSCI, CES, CSRM, DCMD, DITD, ESMD, GEO, POP, TAB, CDF, Econometrica Inc., Galois, Knexus Research Corp, MITRE, NLT, TI, and Tumult Labs.

We also acknowledge and greatly appreciate the ongoing feedback we have received from external stakeholder groups that has contributed to the design and improvement of the Disclosure Avoidance System.

For more information and technical details relating to the issues discussed in these slides, please contact the author at michael.b.hawes@census.gov.

Any opinions and viewpoints expressed in this presentation are the author's own, and do not represent the opinions or viewpoints of the U.S. Census Bureau.

TDA System Requirements

The 2020 Disclosure Avoidance System's TopDown Algorithm (TDA) will implement formal privacy protections for the P. L. 94-171 Redistricting Data Summary File, Demographic Profiles, Demographic and Housing Characteristics, and Special Tabulations of the 2020 Census.

TDA system requirements include:

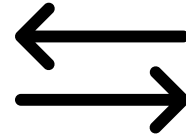
- Input/Output specifications
- Invariants
- Constraints
- Utility/Accuracy for pre-specified tabulations
- ϵ -asymptotic consistency
- Transparency

TDA Process Snapshot



What is a histogram?

Record ID	Block	Race	...	Sex
1	1001	Black	...	Male
2	1001	Black	...	Male
3	1001	Asian	...	Female
4	1001	Asian	...	Female
5	1001	Black	...	Male
6	1001	AIAN	...	Female
7	1001	AIAN	...	Male
8	1001	Black	...	Female
9	1001	Black	...	Female



Attribute Combination (Block/Race/.../Sex)	# of Records
1001/AIAN/.../Male	1
1001/AIAN/.../Female	1
1001/Asian/.../Male	0
1001/Asian/.../Female	2
1001/Black/.../Male	3
1001/Black/.../Female	2
...	...

Histogram: Record count for each unique combination of attributes (including location)

Microdata: One record per respondent

Noisy Measurements

TDA allocates shares of the total privacy-loss budget by geographic level and by query.

Each query of the confidential data will have noise added to its answer.

The noise is taken from a probability distribution with mean=0, and variance determined by the share of the PLB allocated to that particular query at that geographic level.

These noisy measurements are independent of each other, and can include negative values, hence the need for post-processing.



What is noise?

To protect privacy, TDA randomly adds or subtracts a small amount from each statistic it calculates from the confidential data.

Attribute Combination (Block/Race/.../Sex)	# of Records
1001/AIAN/.../Male	1
1001/AIAN/.../Female	1
1001/Asian/.../Male	0
1001/Asian/.../Female	2
1001/Black/.../Male	3
1001/Black/.../Female	2
...	...

Total: $9+0=9$

Male: $4+0=4$

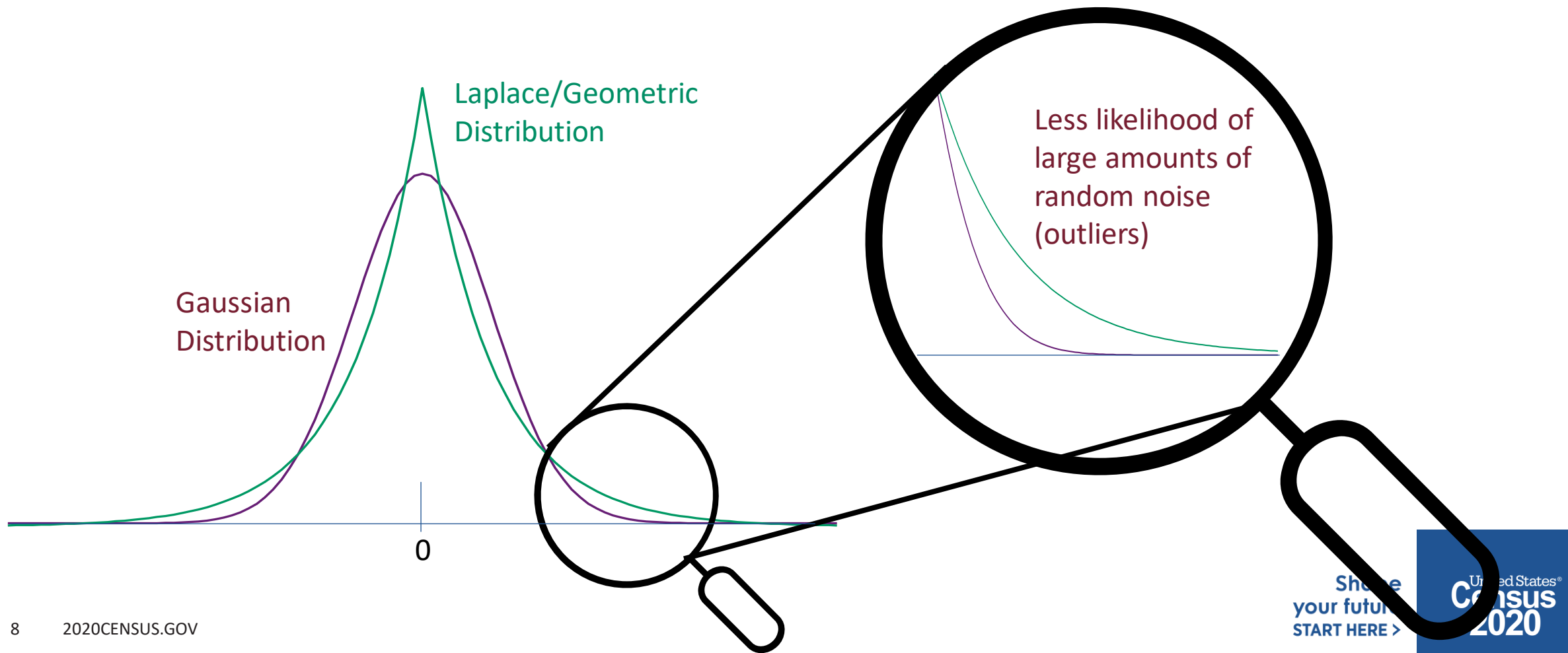
Female: $5-1=4$

#AIAN: $2+0=2$

#Asian: $2+2=4$

#Black: $5-1=4$

Zero-Concentrated Differential Privacy (zCDP)



Understanding *epsilon*, *delta* and *rho*

In traditional $(\epsilon, 0)$ differential privacy:

The privacy-loss parameter ϵ (*epsilon*) sets the upper-bound on how much information leakage can occur.

Shares of ϵ are allocated to each query and sum to the global value of ϵ .

In zero-concentrated differential privacy (zCDP):

Privacy loss is quantified by the paired parameters ϵ and δ (*delta*).

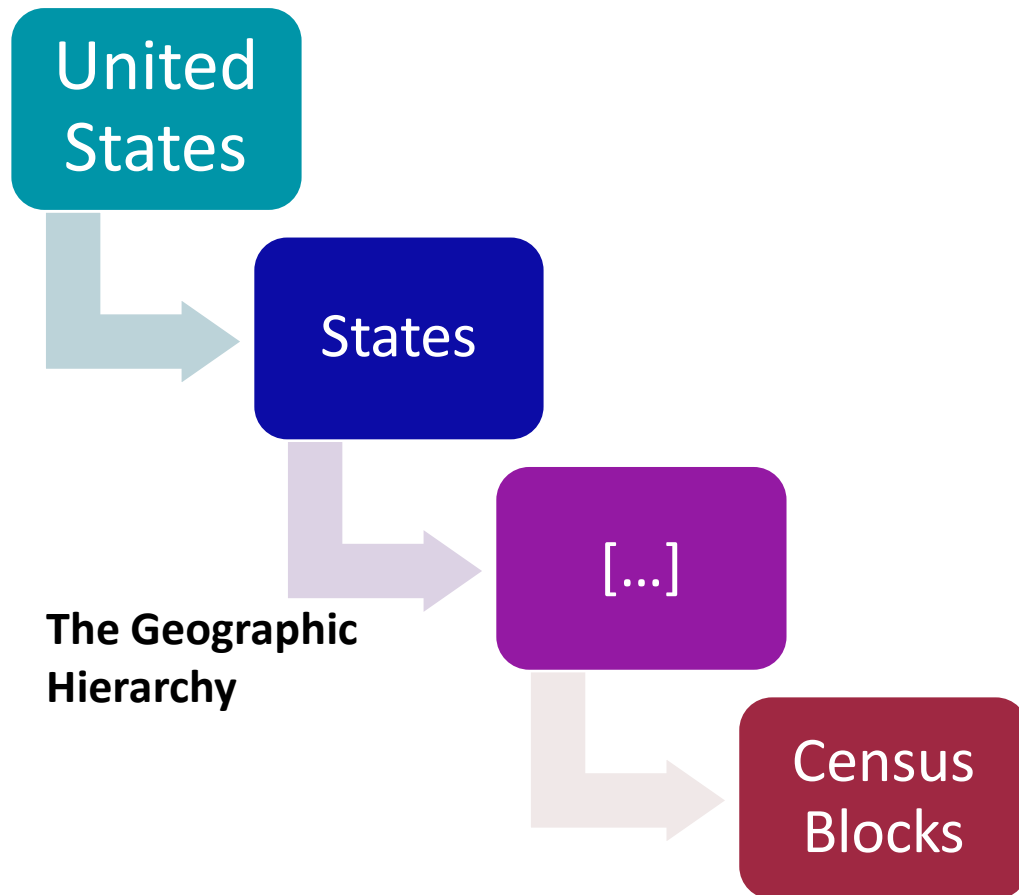
δ is a probabilistic term that establishes the likelihood that privacy loss might exceed the upper bound represented by a particular value of ϵ .

Within the mechanics of zCDP, privacy-loss budget is allocated to queries by shares of a third parameter, ρ (*rho*).

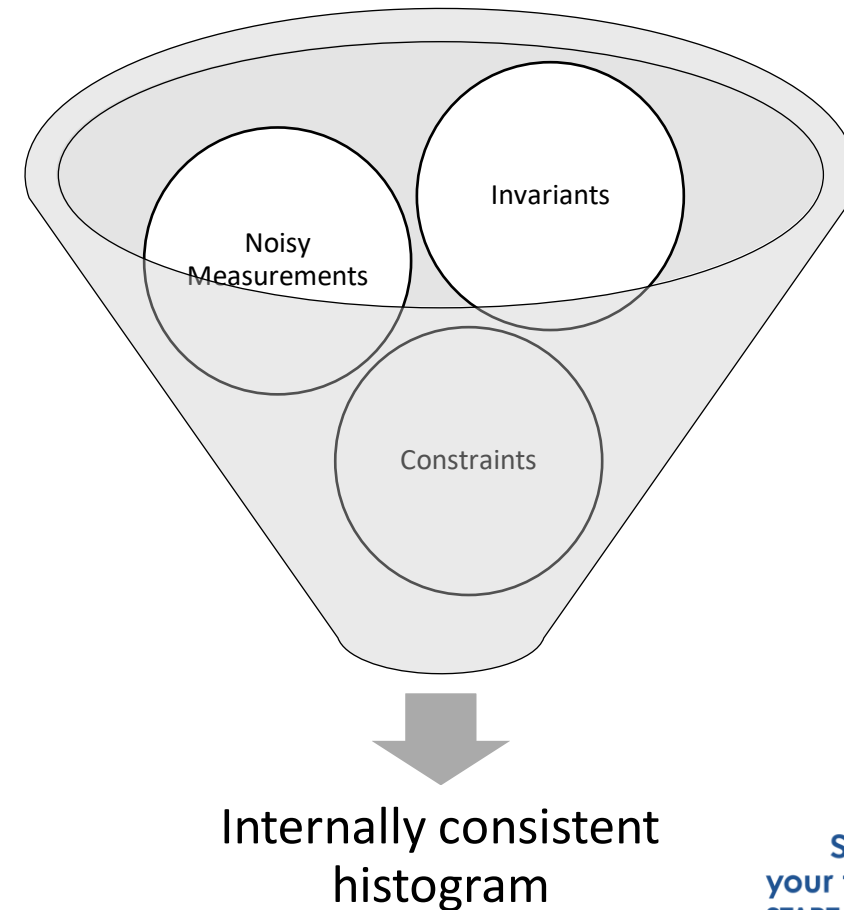
The global ρ can then be used to calculate the global ϵ for any given level of δ .

The Census Bureau's privacy accounting uses a value of $\delta=10^{-10}$ so our published values of ϵ should be interpreted accordingly.

The TopDown Approach



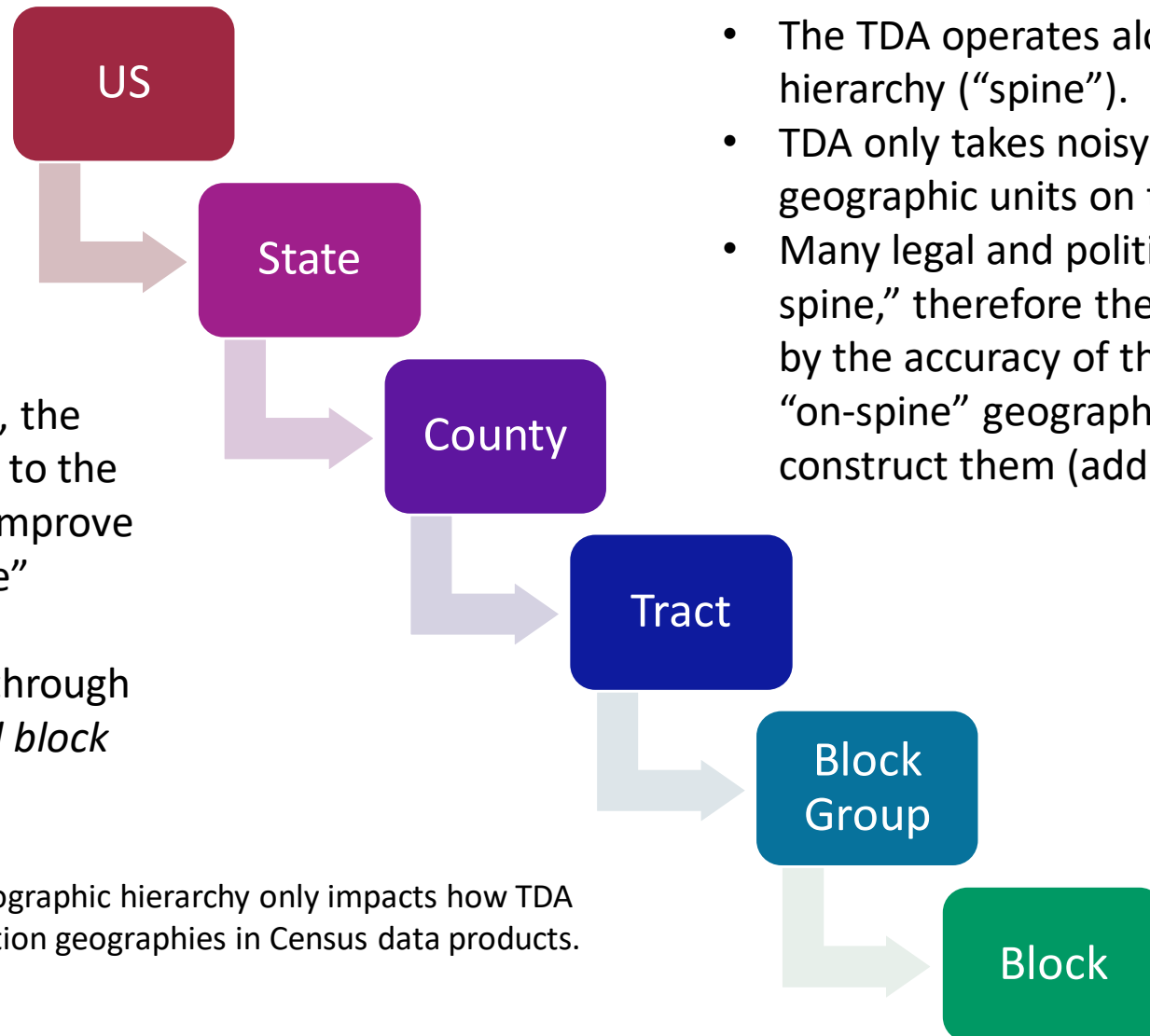
At each geographic level:



Benefits of TDA Compared to Block-by-block

- TDA is in stark contrast with naïve alternatives (e.g., block-by-block or bottom-up)
- TDA disclosure-limitation error does not increase with number of contained Census blocks in the geographic entity
- TDA yields increasing relative accuracy as the population being measured increases (in general), and increased count accuracy compared to block-by-block
- TDA “borrows strength” from upper geographic levels to improve count accuracy at lower geographic levels (e.g., for sparsity)

Tabulation Geographic Hierarchy



- To address this challenge, the DAS Team made changes to the geographic hierarchy to improve the accuracy of “off-spine” geographies.
- This was done primarily through the creation of *optimized block groups (not shown)*.

- The TDA operates along a geographic hierarchy (“spine”).
- TDA only takes noisy measurements for geographic units on the hierarchy.
- Many legal and political geographies are “off-spine,” therefore their accuracy is impacted by the accuracy of the minimum number of “on-spine” geographies that can be used to construct them (adding or subtracting).

Note: The optimization of the geographic hierarchy only impacts how TDA operates. It will not affect tabulation geographies in Census data products.

Rethinking the Geographic Hierarchy

Geographic Hierarchy for Disclosure Avoidance System Processing

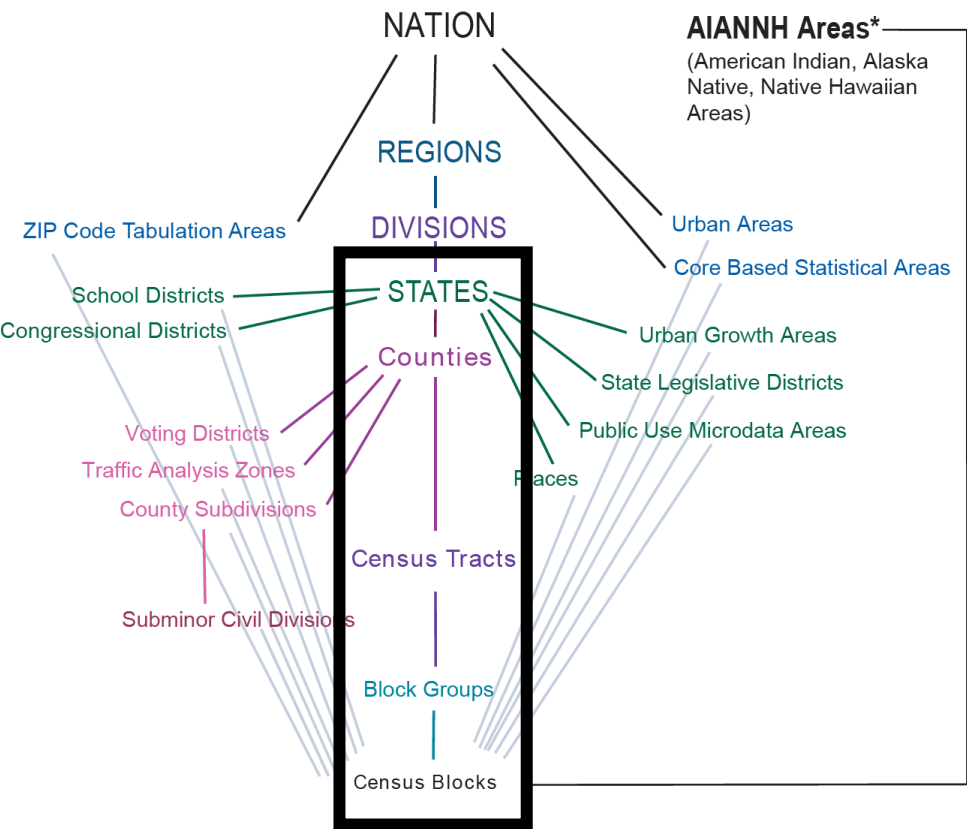
Challenge: Provide for the direct measurement of population and characteristics for American Indian/Alaska Native/Native Hawaiian (AIANNH) areas and sub-state legal geography when applying differential privacy methods.

Consideration: The larger the number of geographic areas on the geographic hierarchy (“spine”) and the more intersections between geographic areas that are formed when one type of area overlaps with another, the more thinly the privacy-loss budget is distributed, impacting the accuracy of data for all geographic areas.

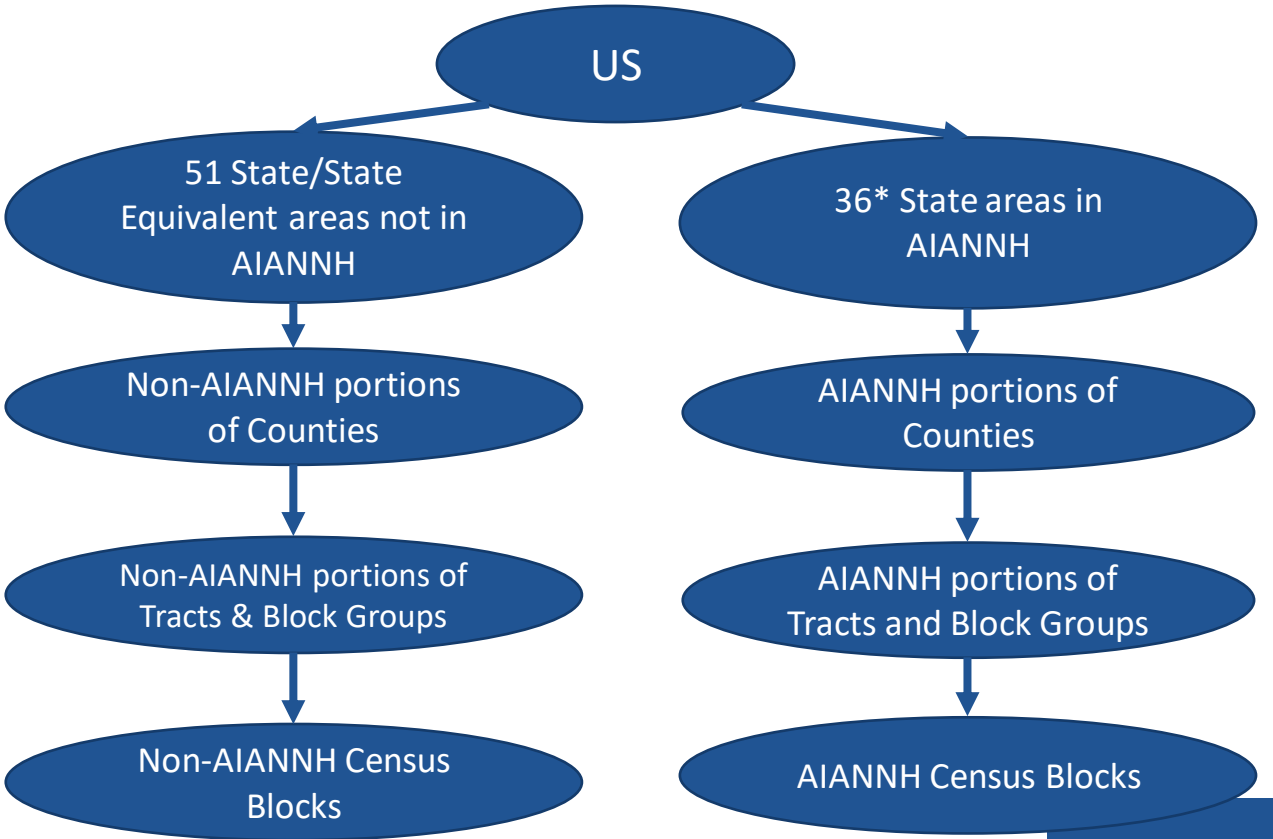
Solution: Bring legal AIANNH areas as well as places (incorporated places and census designated places in 38 states; cities and towns/townships in 12 states) closer to the spine for Disclosure Avoidance System (DAS) processing.

Revising the geographical hierarchy for disclosure avoidance processing

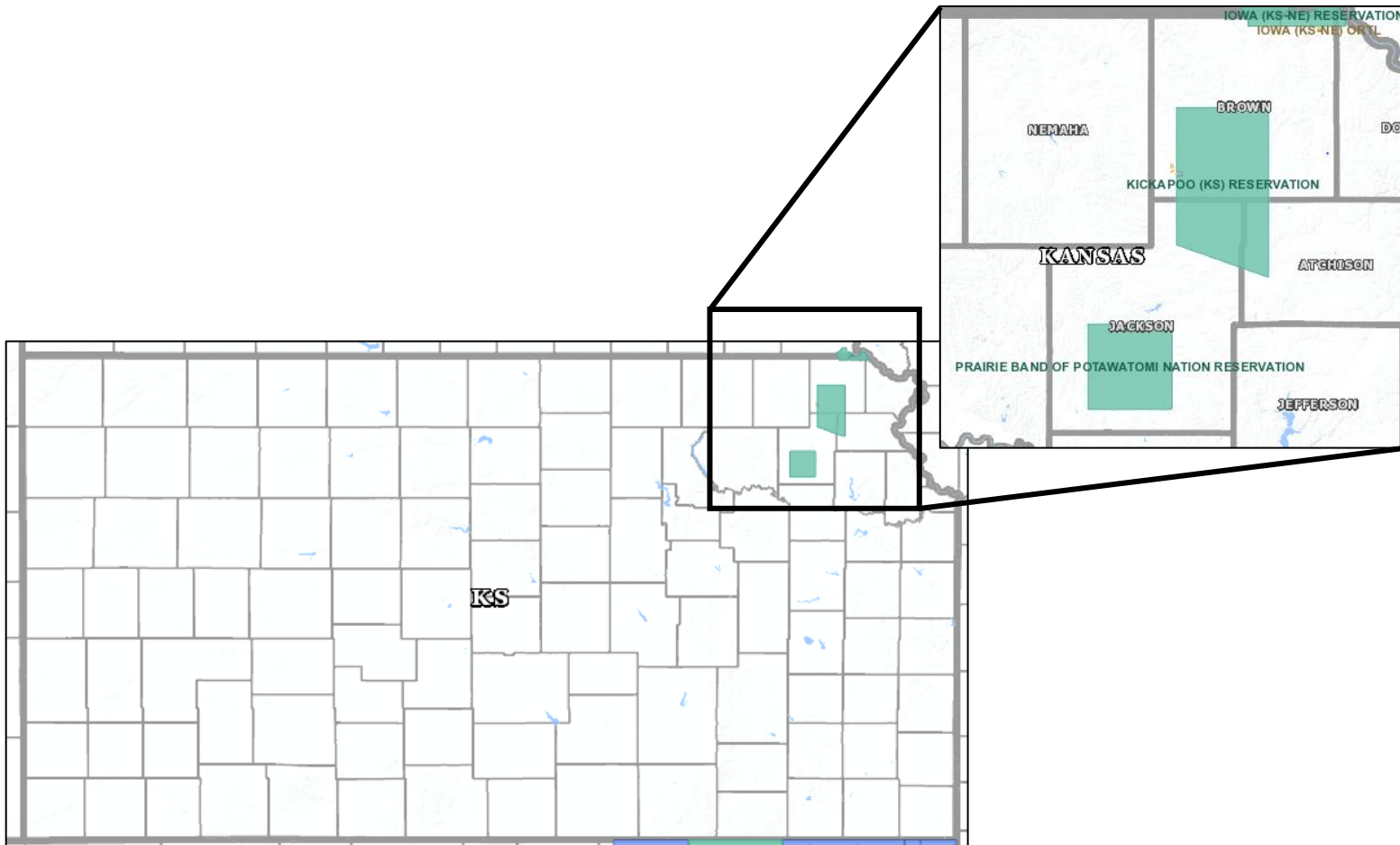
Standard Hierarchy:



Hierarchy for DAS Processing (high-level):



Providing for Direct Measurement of American Indian, Alaska Native, and Native Hawaiian Areas



All AIANNH Areas within the state as a single group, providing a population count for all areas within a state. This minimizes the likelihood that post-processing could result in systematic undercounts.

Example:

The three American Indian areas in Kansas grouped together at the “state” level:

- Iowa (KS-NE) Reservation and Off-Reservation Trust Lands +
- Kickapoo (KS) Reservation +
- Prairie Band of Potawatomi Nation Reservation.

Focusing the geographic hierarchy on the more important sub-state geographic entities in recognition of the regional variations that exist.

Optimized Block Groups (high-level):

In the 38 “non-strong-Minor Civil Division” States, District of Columbia, and Puerto Rico:


Optimized Block Groups were configured to bring Places (Summary Level 160) closer to the spine.



Places

In the 12 “Strong-Minor Civil Division” States:

Optimized Block Groups were configured to bring Minor Civil Divisions (e.g., cities, boroughs, and towns/townships) closer to the spine.



Cities, Boroughs,
and
Towns/Townships

Multi-pass Post-processing

The sparsity of many queries (i.e., prevalence of zeros and small counts) has the potential to introduce bias in TDA's post-processing.

To address the sparsity issue, TDA processing is now performed in a series of passes.

At certain geographic levels, the algorithm constructs histograms for a subset of queries in a series of passes for that level, constraining the histogram for each pass to be consistent with the histogram produced in the prior pass.

Example for the P.L. 94-171 Redistricting Data Summary File:

Pass 1: Total Population

Pass 2: Remaining tabulations supporting P.L. 94-171 Redistricting Data

Sample Privacy-loss Budget Allocation (by geographic level)

Privacy-loss Budget Allocation April 28, 2021				
PPMF				
Person Tables (PPMF-P)				
United States				
	Global rho	192721/184041 (1.05)		
	Global epsilon	10.3		
	delta	10 ⁻¹⁰		
		rho Allocation by Geographic Level		
	US	51/1024		
	State	153/1024		
	County	78/1024		
	Tract	51/1024		
	Optimized Block Group*	172/1024		
	Block	519/1024		

Privacy-loss Budget Allocation April 28, 2021				
PPMF				
Units Tables (PPMF-U)				
United States				
	Global rho	919681/20241001 (0.045)		
	Global epsilon	1.9		
	delta	10 ⁻¹⁰		
		rho Allocation by Geographic Level		
	US	1/1024		
	State	1/1024		
	County	18/1024		
	Tract	75/1024		
	Optimized Block Group*	906/1024		
	Block	23/1024		

Sample Privacy-loss Budget Allocation (by query)

Query	Per Query rho Allocation by Geographic Level					
	US	State	County	Tract	Optimized Block Group*	Block
TOTAL (1 cell)		678/1024**	342/1024	1/1024	572/1024	1/1024
CENRACE (63 cells)	2/1024	1/1024	1/1024	2/1024	1/1024	2/1024
HISPANIC (2 cells)	1/1024	1/1024	1/1024	1/1024	1/1024	1/1024
VOTINGAGE (2 cells)	1/1024	1/1024	1/1024	1/1024	1/1024	1/1024
HHINSTLEVELS (3 cells)	1/1024	1/1024	1/1024	1/1024	1/1024	1/1024
HHGQ (8 cells)	1/1024	1/1024	1/1024	1/1024	1/1024	1/1024
HISPANIC*CENRACE (126 cells)	5/1024	2/1024	3/1024	5/1024	3/1024	5/1024
VOTINGAGE*CENRACE (126 cells)	5/1024	2/1024	3/1024	5/1024	3/1024	5/1024
VOTINGAGE*HISPANIC (4 cells)	1/1024	1/1024	1/1024	1/1024	1/1024	1/1024
VOTINGAGE*HISPANIC*CENRACE (252 cells)	17/1024	6/1024	11/1024	17/1024	8/1024	17/1024
HHGQ*VOTINGAGE* HISPANIC*CENRACE (2,016 cells)	990/1024	330/1024	659/1024	989/1024	432/1024	989/1024

*The Optimized Block Groups used within the TopDown Algorithm differ from tabulation block groups. These differences improve accuracy for "off-spine" geographies like places and minor civil divisions. The use of optimized block groups for measurement and post-processing within the TopDown Algorithm does not impact how the resulting data will be tabulated. All Census data products will be tabulated using the official tabulation block groups as defined by the Census Bureau's Geography Division.

**The TOTAL query (total population) is held invariant at the state level. This rho allocation assigned to TOTAL at the state level is the amount assigned to the state-level queries for the total population of all American Indian and Alaska Native (AIAN) tribal areas within the state and for the total population of the remainder of the state, for the 36 states that include AIAN tribal areas.

Webinar Series:

Understanding the 2020 Census Disclosure Avoidance System

All webinars start at **1:00 pm EDT**

No pre-registration necessary.

*Search “*disclosure webinars*” at www.census.gov for log-in information and archived presentations.

Or go to: <https://www.census.gov/data/academy/webinars/2021/disclosure-avoidance-series.html>




Day	Date	Title
T	May 4	Differential Privacy 101
F	May 7	The Census Bureau's Simulated Reconstruction-Abetted Re-identification Attack on the 2010 Census
Th	May 13	Differential Privacy 201 and the TopDown Algorithm
F	May 14	Highlights of the April 2021 Detailed Summary Metrics
F	May 21	Analysis of April 2021 Demonstration Data for Redistricting and Voting Rights Act Use Cases

Shape
your future
START HERE >

United States[®]
Census
2020

Stay Informed:
Subscribe to the 2020 Census Data
Products Newsletters

*Search “Disclosure Avoidance” at www.census.gov



2020 Census Data Products Newsletters

Sign up for news and information about 2020 Census Data Products and the implementation of the new Disclosure Avoidance System.

SIGN-UP FOR NEWSLETTERS

Past Issues:

May 04, 2021
Webinar Today (5/4): Differential Privacy 101

April 30, 2021
Save the Dates for Additional Webinars Throughout May

April 28, 2021
New DAS Update Meets or Exceeds Redistricting Accuracy Targets

April 19, 2021
New Demonstration Data Will Feature Higher Privacy-loss Budget

April 07, 2021
Meeting Redistricting Data Requirements: Accuracy Targets

February 23, 2021
The Road Ahead: Upcoming Disclosure Avoidance System Milestones

Stay Informed: Visit Our Website

*Search “Disclosure Avoidance” at www.census.gov

***“Disclosure Avoidance Webinar Series:
Join live or view archived presentations”***

2020 Census Data Products: Disclosure Avoidance Modernization

Modern computers and today's data-rich world have rendered the Census Bureau's traditional confidentiality protection methods obsolete. Those legacy methods are no match for hackers aiming to piece together the identities of the people and businesses behind published data.

A powerful new disclosure avoidance system (DAS) designed to withstand modern re-identification threats will protect 2020 Census data products (other than the apportionment data; those state-level totals remain unaltered by statistical noise).

Inspired by cryptographic principles, the 2020 DAS is the only solution that can respond to this threat while maximizing the availability and utility of published census data.



Protecting Privacy with Math

Learn More:

- ** Disclosure Avoidance Webinar Series: Join live or view archived presentations **
- Census Bureau Declarations for Alabama v. Commerce II Litigation [4.2 MB]
- Video Presentation: Differential Privacy and the 2020 Census [242 MB]
- Animation: Protecting Privacy with Math, a collaboration with MinutePhysics
- Infographic: A History of Census Privacy Protections
- JASON report on Privacy Methods for the 2020 Census
- All Disclosure Avoidance Working Papers



Census Privacy Protection History

Latest Updates

- Disclosure Avoidance System Development

Data Products Newsletter

April 30, 2021

Save the Dates for Additional Webinars Throughout May

[SIGN-UP FOR NEWSLETTERS](#)

[VIEW ALL NEWSLETTERS](#)